The One-boxing Intuition In Newcomb's Problem

Examination Number: 4158847

Philosophy Taught Masters

The University of Edinburgh

2010

Contents Page

Contents Page

The One-boxing Intuition in Newcomb's Problem

ABSTRACT: In Newcomb's Problem, a lot of people choose the option of one-boxing. However, the equivalent choice is rarely made in other similar decision problems. This discrepancy needs explaining. In this thesis, I consider Horwich's suggestion that evidential decision theory provides an explanation. Horwich argues that evidential decision theorists make these different choices because static screening cannot be used in Newcomb's Problem, yet it can be used in other similar decision problems. I discuss and expand upon his argument.

I then describe the process of dynamic screening. The appeal to dynamic screening purports to show how screening could occur in Newcomb's Problem and thus lead evidential decision theorists to two-box. However, I argue that dynamic screening would not convince an evidential decision theorist to two-box, since the argument behind dynamic screening presupposes the irrationality of evidential decision theory. I then argue that using static screening also presupposes the irrationality of evidential decision theory. Thus, Horwich's argument is flawed. Screening cannot occur in any common cause problems, and the evidential decision theorist would be expected to do the equivalent of one-boxing. Evidential decision theory cannot explain why agents do the equivalent of two-boxing in some decision problems, and yet one-box in Newcomb's Problem.

Keywords: Newcomb's Problem, evidential decision theory, screening, common-cause, one-boxing.

Section 1: Introduction

1) Introduction

Some people choose to one-box in Newcomb's problem (NP), yet no-one chooses the equivalent option in other similar problems. In this thesis, I consider whether evidential decision theory (EDT) can explain this difference. NP can be stated as follows:

'The Predictor is a being who is able to predict your choices with great accuracy... There are two boxes, a transparent box b1, and an opaque box b2. You can see that b1 contains $1000; b2 contains either $1,000,000 or nothing...either you take what is in both boxes, or you take what is the opaque box b2 alone. If the Predictor predicted that you will take what is in both boxes, he does not put $1,000,000 in b2; but if the Predictor predicted that you will take what is in b2 alone, he puts $1,000,000 in b2. You value more money to less money. What should you do?'[1]

In making a decision, two lines of reasoning give conflicting advice:[2]

1. If you one-box, then the Predictor is likely to have predicted this, meaning you receive $1million. If you two-box, you expect the Predictor to have predicted this, leaving you with $1000. You should therefore one-box.
2. The Predictor has already put $1million or nothing in b2. Whichever obtains, two-boxing is better because you receive $1000 more. You should therefore two-box.

The first line of reasoning uses EDT, which calculates expected value based on

1 Ninan pp1, from Nozick (1969).
2 Ninan pp1, Nozick (1993) pp42, Nozick (1969) pp115.

the evidence that your actions provide for various outcomes. The second employs causal decision theory (CDT), which calculates expected value by considering causal consequences of actions.[3]

In this thesis I begin by discussing several cases (hereafter 'the other common cause cases') which offer the equivalent of one-boxing and two-boxing choices. In these other cases, one-boxing is seen to seem irrational. In contrast, in NP, many people[4] think that it is rational to one-box. This thesis seeks an explanation for the one-boxing intuition (OBI) being present in NP, yet not in the other common cause cases.

I consider Horwich's claim that EDT recommends one-boxing in NP, but in fact recommends 'two-boxing' in the other common cause cases, and thus that EDT is responsible for OBI. The reason that Horwich gives for this difference in the recommendations of EDT is that in the other common cause cases static screening, described by Eells, can seemingly be used. Screening prevents either action being probabilistically relevant to which state[5] obtains, and thus means that arguments like Argument 1 above no longer apply in these other common cause cases. However, Horwich claims that static screening cannot be employed in NP, and thus that whether we one-box or two-box in NP does provide evidence for the Predictor's prediction, as described by Argument 1.

Eells claims that Horwich is wrong in believing that screening cannot occur in NP. Eells proposes that 'dynamic screening' should occur in NP, thus leading EDTers[6] to two-box. I argue that 'dynamic screening' would not convince an EDTer to two-box because it presupposes the irrationality of EDT. If Horwich's arguments are accepted it appears at this point that EDT could explain OBI; static screening can be used in the other common cause cases, but there is no way using static screening in NP. However, I then argue that 'static screening', like dynamic screening, presupposes the irrationality of EDT. Thus, Horwich and Eells are both mistaken in supposing that the EDTer can use static screening to reach the 'two-boxing'[7] conclusion in the other common cause cases.

3 Joyce pp3. For how EV is calculated according to both decision rules is see Burgess pp263-4, pp276.
4 Nozick (1997) pp48 claims around 50%.
5 E.g whether there is $1million in box b2 or not.
6 Agents that use evidential decision theory.
7 I use 'one-boxing' and 'two-boxing' to refer to the actions in the other common cause cases that are the equivalent to one-boxing and two-boxing in NP.

Whilst EDTers should one-box in NP, they should also 'one-box'[8] in the other common cause cases. Therefore, EDT cannot explain why OBI is present in NP but not in other cases.

2) The structure of the problem

I now explain in fuller detail the structure of NP, in order to clarify probabilistic references used throughout this thesis. NP is a 'common cause' problem.[9] Since the Predictor's prediction is highly indicative of the agent's action, but does not cause it, a 'common cause' of the two events is postulated.[10] For example, this common cause could be a feature of the agent's brain that has two different appearances, one of which causes one-boxing and the other which causes two-boxing. The Predictor analyses this feature to make his prediction. Thus the problem's structure can be represented as in the following diagram, in which the arrows represent causation:

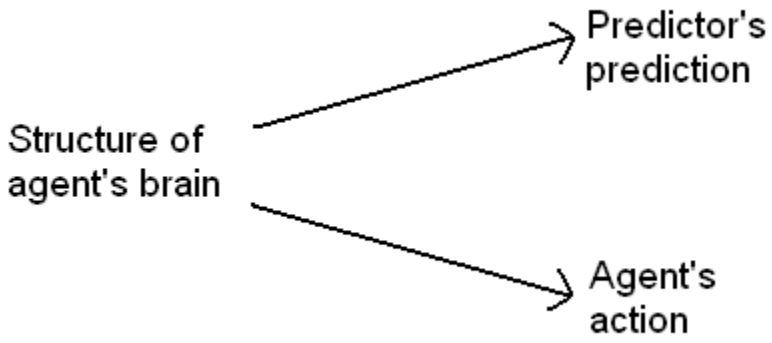


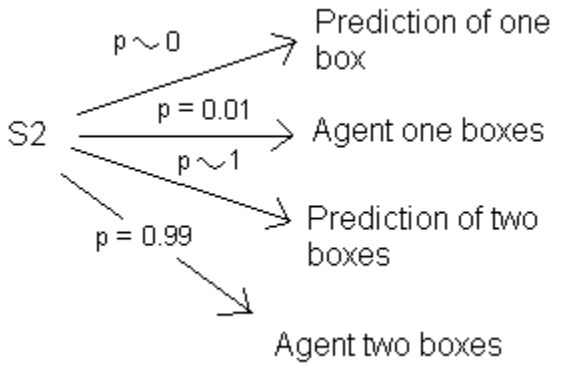
Figure 1: The common cause structure of NP.

8 The quotes illustrate that the action in question is not in fact one-boxing, but rather is the equivalent action in other cases that have the same structure as NP.

9 Eells (1982) analysed the common cause structure of NP, discussed in Burgess.

10 We could postulate that the correlation is pure co-incidence, but this is not what we would expect of a rational agent (Eells (1989) pp12)

The Predictor's accuracy is taken to be 99% in most discussions of NP. This means that the probability of the agent two-boxing, given that the Predictor has predicted he will two-box, is 0.99 and the probability of the agent one-boxing given that the Predictor has predicted he will one-box is also 0.99: P (agent two-boxing | prediction of two-boxing) = 0.99 and P (agent one-boxing | prediction of one-boxing) = 0.99. This can be illustrated with the following probabilistic structures, in which 'S2' and 'S1' are states of the agent's brain that predispose to two-boxing and to one-boxing, respectively:



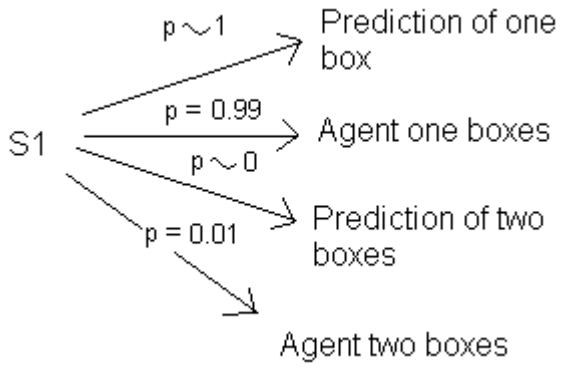


Figure 2: The probabilistic common cause structure of NP.

This diagram illustrates that if the agent has a particular brain structure, either of

a one-boxer (S1) or a two-boxer (S2), then the Predictor will make a prediction that corresponds 100% of the time to the agent's brain structure, while the agent will act in accordance with his brain structure 99% of the time.

Argument 1) in sub-section 1 above relies on the premise, 'if you one-box, then the Predictor is highly likely to have predicted this.' This 'high likelihood' is frequently referred to in the literature as 99%.[11] However, this inverse probability, relied upon in Argument 1, is not actually stated in any descriptions of the problem. All we are told is that, if the agent predicts that you will one-box, then the probability that you will one-box is 99%.

I now consider the implications of inferring that P (prediction of two-boxing | agent two-boxing) = 0.99 from the given statement that P (agent two-boxing | prediction of two-boxing) = 0.99.

Bayes theorem claims that: $P(A \mid B) = (P(B \mid A)\, P(A)) \div P(B)$.

Thus, the probability of the Predictor having predicted it if you one-box is only equal to the probability of you one-boxing if the Predictor predicts it, if the probability of the Predictor predicting you one-box is the same as the probability of you one-boxing. For instance, consider what would result if the population were divided up as follows:[12]

11 For example Burgess (2004).
12 There is a similar example in Levi (1975) pp164-166.

	Predicts one-box	Predicts two-box
One-boxes	98.01%	0.01%
Two-boxes	0.99%	0.99%

Figure 3: Hypothetical population probabilities.

In this case, P (agent two-boxes| prediction of two-boxing) = 99%

However, P (prediction of two-boxing| agent two-boxes) = 50%

In order to infer from P (agent two-boxes | prediction of two-boxing) = 99% that P (prediction of two-boxing | agent two-boxes) = 99% the following unique structure, derived in Appendix A, must be assumed:

	Predicts one-box	Predicts two-box
One-boxes	49.50%	0.50%
Two-boxes	0.50%	49.50%

Figure 4: Population probabilities required for assuming that P (prediction of two-boxing| agent two-boxes) = 99% and P (agent two-boxes/prediction of two-boxing) = 99%.

However, for argument 1) to hold, P (prediction of one-boxing | agent one-boxes) does not in fact need to be as high as 99%. I show in Appendix B that it need only be greater than 50.05% in order for EDT to recommend one-boxing. Thus, there is in fact considerably more flexibility in the structure of the problem than suggested in Figure 4 (discussed mathematically in Appendix C).

Thus, I have described the structure of NP. Argument 1 would place severe restrictions on the population make-up if P (Predictor predicts that you one-box | you one-box) is specified to be 99%. However, I have shown that, for Argument 1 to hold, P

(Predictor predicts that you one-box | you one-box) need not be as high as 0.99, but only as high as 0.5005. Thus the make-up of the population is not as restricted as it first appears.

Section 2: The Static Defence (Other common cause cases)

3) Common cause cases

I now describe three cases which have the equivalent of 'one-boxing' and 'two-boxing' actions.' As with NP, they are all common cause cases.

In the Jones-ruthlessness case[13], Jones competes for a promotion to be decided by psychological testing. After the test, Jones learns that whoever scored highest on ruthlessness in the test will be promoted. He then has to decide whether to fire Smith. Jones believes it would be best for the company not to fire Smith, but he very much wants the promotion. Should Jones fire Smith? Firing him (the equivalent of one-boxing) would be evidence for ruthlessness. Yet it certainly appears irrational to fire Smith in order to receive a good 'news-value'[14] concerning one's ruthlessness, because one's promotion has been predetermined by the test.

The Jones-ruthlessness case has the same structure as NP. NP has the following structure, where 4, 3,2, and 1 represent the most-preferred to the least-preferred pay-off to the agent:

	One-boxes	Two-boxes
Agent has one-boxing common cause	**3**	4
Agent has two-boxing common cause	1	**2**

Figure 5: The structure of NP.

13 Hurley (1991) pp175, Nozick (1969) pp135.
14 Joyce's terminology.

There are two important features of NP, that are embodied in the arguments of the EDTer and CDTer. Firstly, one-boxing gives either pay-offs of 1 or 3, whereas two-boxing gives pay-offs of 2 or 4. Thus, two-boxing strictly dominates one-boxing, since, whatever common cause the agent has, he does better by two-boxing. Secondly, the high correlations between common cause and action mean that the two emboldened combinations (2 and 3) are the most likely to obtain. Of these, the agent would prefer to one-box for a pay-off of 3 than two-box for a pay-off of 2. The conflict between these two considerations creates NP.

The Jones-ruthlessness case has the same structure as NP, once it is stipulated that getting the job is much more important to the agent than not firing Smith.[15]

	Fires Smith	Doesn't fire Smith
Ruthless	**3**	4
Not ruthless	1	**2**

Figure 6: The structure of the Jones ruthlessness case.

The second case for consideration is the Smoker's Fantasy. In the Smoker's Fantasy[16] the agent is deciding whether to smoke. He gets pleasure from smoking, but he also knows that there is a strong correlation between smoking and lung cancer. He believes that this correlation is explained by, 'a powerful genetic common cause of both smoking and cancer'.[17] Should he smoke? Again, once it is stipulated that it is much more important to the agent that he avoids cancer than that he gets the pleasure from smoking, the structure of the case is identical to NP. In this case EDT recommends abstaining, since smoking provides strong evidence for having the cancer-causing gene. Again, however, abstaining seems irrational. Whether or not you have the gene, you would rather be getting pleasure from smoking, and your action cannot causally influence which pre-existing state obtains.

15 These pay-offs are the equivalent of the $1m and $1000 in NP
16 Eells (1984) pp73
17 Ibid.

The third case is the athlete vs academic case, in which P is deciding whether to become an academic or a basketball player. P does not know whether his father is S or T, and he is concerned since S died of 'some terrible inherited disease'[18], which is genetically dominant. Furthermore, it is known that the 'tendency to decide to do acts which form part of an intellectual life'[19] is genetically dominant, and that S had this tendency. P prefers the life of an academic to that of an athlete. Again, the structure of this problem is identical to NP, since the agent would rather avoid the terrible disease than fulfil his desire of becoming an academic. EDT would in this case recommend becoming a basketball player, since this would provide evidence that P will not get the terrible disease. Yet, again this seems irrational. The job that P chooses cannot causally influence which man is his father. Whoever P's father is, it would be better to pursue job fulfillment.

Thus, our intuition does not accord with EDT's recommendations in these cases. EDT appears to be just, 'a futile attempt to manipulate the cause by suppressing its symptoms'.[20] If EDT is in fact responsible for OBI, a difference between Newcomb's Problem and these cases must be highlighted that explains why the recommendations of EDT are taken seriously in NP.

4) Screening

Horwich claims that EDT could explain OBI. Such a theory needs to illustrate how EDT can recommend 'two-boxing' in the other common cause cases, whilst recommending one boxing in NP. So far, it seems that EDT recommends 'one-boxing' in the other common cause cases, and that this recommendation seems irrational. In this thesis, I do not intend to address the question of whether EDT is in fact irrational by virtue of its prescribing actions that provide evidence for states that it could have no causal influence over. What I intend to address is whether, if there are people who follow the prescriptions of EDT (whether irrational, or not), this can explain OBI. Yet,

18 Nozick (1997) pp 56.
19 Ibid.
20 Skyrms (1980) pp129.

since OBI is not present in the other common cause cases, if EDT did recommend one-boxing in these cases, this would mean that EDT could not explain OBI. The current situation is illustrated in this figure:

	EDT recommends	Do we have an intuition to one-box?
Newcomb's Problem	One-boxing	Yes
Other common cause cases	'One-boxing'	No, 'one-boxing' seems absurd.

Figure 7: An illustration of the problem with EDT explaining OBI.

However, Eells argues that in the other common cause cases EDT actually recommends 'two-boxing,' and thus that our intuitions in these cases are in line with what EDT recommends. Eells claims that static screening can be used in these cases, which causes EDTers to 'two-box.' In fact, Eells believes that screening can occur in all common cause cases (I return to this in sub-section 17). On the other hand, Horwich argues that screening cannot occur in NP since it, 'works in normal circumstances...but..no such strategy is available in the very special conditions of NP.'[21] This explanation is illustrated in Figure 8:

21 Horwich (1985) pp435.

	EDT recommends	Do we have an intuition to one-box?	Can screening occur?
Newcomb's Problem	One-boxing	Yes	No
Other common cause cases	'Two-boxing'	No	Yes

Figure 8: An illustration of Horwich's potential explanation for OBI.

If the information in Figure 8 were true, EDT could explain OBI. Horwich's argument that EDT could explain OBI must defend two claims:

a) Static screening can be used in other common cause cases, with the result that EDT recommends two-boxing.

b) Static screening cannot be used in NP.

5) Claim a)

I begin by considering how screening[22] in the other common cause cases leads to EDT recommending 'two-boxing'. Eells first described the process of static screening in order to show how EDT and CDT in fact converge in their recommendations. Eells's description of static screening, using the Smoker's Fantasy, is as follows:[23]

1) The agent is rational in the Bayesian sense, and knows this. (I return to consider the truth of this assumption in sub-sections 12 and 14).

22 Reichenbachs terminology (Eells and Sober (1986) pp228)
23 Eells (1984) pp73-76.

2) The rational agent will, and knows that he will, base his action on just three things: his beliefs (subjective probabilities), his desires (desirabilities), and his deliberation in the light of his beliefs and desires (the application of some principle of expectation maximisation).

3) If the common cause is to affect his action, it must affect one or more of these three things.

4) The common cause does not affect his deliberation i.e. which decision rule he will use. (I return to consider truth of this assumption in sub-sections 10 and 14).

5) The rational agent is aware of 3) and 4).

6) Thus, the rational agent believes that the gene causes rational people like himself to smoke by causing them to have beliefs and desires such that a rational evaluation of the available acts in light of those beliefs and desires leads to the conclusion that smoking is rational (from 1, 2, 3, 4 and 5).

7) The agent knows his own subjective probabilities and desirabilities (I return to consider this the truth of this assumption in sub-section 6).

8) Thus, given that the agent holds certain beliefs and desires, his probability of smoking is independent of whether he has the smokers' gene or not. His beliefs and desires alone will affect his action; rendering the gene's presence irrelevant. The following equality illustrates this:

$P (S \mid R \& G) = P (S \mid R \& -G)$ where S= smoking, G= having the smoking gene, R = the agent's beliefs and desires.

9) R is the proposition that attributes to the agent the subjective probabilities and desirabilities that he does in fact have, therefore $P (R) = 1$. Thus, given the subjective probabilities and desirabilities that the agent has, it is true that:

$P (S \mid G) = P (S \mid -G)$

10) By symmetry of probabilistic independence:

$P (G \mid S) = Pr(G \mid -S)$.

Thus , smoking does not change the probability of having the smokers' gene.

11) Given this equality, since the agent has a desire to smoke, it follows that EDT, like CDT, will recommend smoking.

Essentially, Eells's appeal to static screening shows that EDT cannot recommend an action based on that action providing evidential relevance, since the agent should, by introspection, be aware of which state he is in before he acts. The agent's awareness of his state prevents his actions having any probabilistic relevance to the previously obtaining states. Thus, as Eells intended to illustrate, both EDT and CDT should recommend smoking in the Smoker's Fantasy.

I now turn to consider whether this could feasibly provide an explanation for OBI by considering why Horwich believes that static screening cannot be used in NP (Claim b):

6) Claim b)

I now turn to consider Horwich's suggestion for why screening could not occur in NP. Horwich claims that, in cases such as NP, the agent does not have enough information to screen. Thus, while EDT recommends two-boxing in the other common cause problems, it recommends one-boxing in NP.

What then is the information that Horwich believes the agent is missing that prevents him screening? The agent is aware of his beliefs and desires, but lacks awareness of which common cause these beliefs and desires stem from, 'there is something crucial about his beliefs and desires that he doesn't know: namely...whether they are of the kind caused by the common cause or not.'[24] Therefore, at this point, the acts will still have evidential relevance. Formally, Horwich disputes that P (S | R & G) = P (S | R & -G) (Premise 9) holds in NP. As we have seen, this equality claims that if we know our beliefs and desires, then knowing which common cause is present does not change the probability of smoking, since the beliefs and desires alone mediate the

24 Eells (1984) pp87.

agent's action. Empirically, it is a true statement; given the agent's beliefs and desires, the probability of his performing either act will remain constant irrespective of which common cause is present. However, in NP, claims Horwich, knowing which common cause is present would change the agent's *subjective* probability[25] of two-boxing, since it would inform the agent of the common cause to which his beliefs and desires correspond. The agent's subjective probabilities are those used by the Bayesian rational agent in deliberation. By the same reasoning, the performance of an action would provide the rational agent with evidence for which common cause his beliefs and desires stemmed from.[26]

Horwich believes that NP is a case in which screening cannot occur because, 'in Newcomb's problem the causal chain leading to a given act is not mediated by any particular, easily recognisable, introspectible state.'[27] Thus, in NP we would not know if our own beliefs and desires were of the kind that lead to one-boxing or two-boxing. Knowing whether we had the brain of a one-boxer or two-boxer would therefore change our subjective probabilities of a certain action being performed, since we would then know from what common cause our beliefs and desires stem. By the same reasoning, and more importantly, two-boxing or one-boxing would change our assessment of the subjective probabilities of having each common cause. In NP, an EDTer would therefore one-box. Thus Horwich argues that, 'what makes Newcomb's problem bizarre [is] the existence of evidential implications of an act that are not screened off by the agent's awareness of his own motivational state.[28]'

7) Deciphering Nozick

Nozick also suggests that screening is unavailable in NP. His does this by illustrating why static screening can be used in the athlete vs academic case, and claiming that NP does not possess the same features that allow the use of static

25 Horwich (1985) pp440, Sobel and Eells (1986) pp226-227 draw this distinction between first person beliefs and general beliefs.
26 Sober and Eells (1986) pp233 also discuss such a case.
27 Horwich (1985) pp438.
28 Ibid pp435-436.

screening.

Nozick notes that, 'preferring the academic life to the athlete's life should be as strong evidence for the tendency [to become an academic] as choosing the academic life. And hence P's choosing the athlete's life, though he prefers the academic life, on expected utility grounds, does not seem to make it likely that he does not have the tendency'.[29] What Nozick has shown here is that static screening can be used in the athlete vs academic case. It seemed that EDT would recommend becoming a basketball player in this case (sub-section 3). However, once we note that, 'preferring the academic life to the athlete's life should be as strong evidence for the tendency as choosing the academic life,'[30] then it becomes obvious that static screening can be used. P's preference to become an academic informs him that he has inherited the tendency for academia and thus that he will end up with the terrible disease. The action of becoming a basketball player would no longer have positive evidential relevance to the preferred common cause, 'P's choosing the athlete's life..does not seem to make it likely that he does not have the tendency.'[31] Knowing that he will inherit the terrible disease, even the EDTer would seek job fulfillment.

Nozick therefore shows that, in the academic vs athlete case, the agent's preference for the academic life allows the use of static screening. Horwich discusses how the other common cause cases that I considered in sub-section 3 have the same feature. In the Jones ruthlessness case, we are told that Jones believes that Smith doesn't deserve to be fired.[32] This would suggest that Jones is not going to fire Smith and therefore that Jones is not ruthless, 'suppose Jones knows perfectly well that he is not inclined to fire Smith, and that under normal circumstances he would certainly not act in such a ruthless manner. On this basis, he infers that he probably did not do very well on the ruthlessness test.'[33] Given this, it is pointless for Jones to fire Smith to gain evidence that he is ruthless, because he won't be promoted anyway.

In the Smoker's Fantasy, the agent wants to smoke. This provides evidence that

29 Nozick (1997) pp 348.
30 ibid
31 ibid
32 Hurley (1991) pp182.
33 Horwich (1985) pp434.

he will smoke, and thus that he has the smoking common cause and will get cancer. As Horwich says, 'there is a reasonable inference from the statistical data: namely, that there is a high correlation between cancer and having the inclination to smoke'[34] since, 'it is assumed that the gene tends to cause the inclination to smoke, which tends to cause smoking.'[35] You therefore know that you will get cancer, and even the EDTer would smoke to gain the pleasure of smoking, 'smoking is not symptomatic of the bad gene; and so the evidential principle does not dictate that you should abstain.'[36] Not-smoking would not provide any evidence that you will not get cancer, 'your chances of getting cancer will not be diminished in the slightest if you finally decide not to smoke.'[37]

Nozick, like Horwich, does not think that static screening can be used in NP. Nozick claims that, in order to make the athlete-academic case a fair comparison to NP, 'what the example seems to require is an inherited tendency to decide to do A which is such that...the probability of its presence *cannot be estimated on the basis of the person's preference*s (emphasis added), but only on the basis of knowing the genetic make up of his parents, or knowing his actual decisions.'[38] Thus, like Horwich, Nozick believes that in NP we cannot know purely on the basis of our preferences which common cause we have, and thus that the agent's decisions will provide him with evidence for which common cause he possesses. Whilst in the other common cause cases the agent prefers the academic life, prefers not to fire Smith, and prefers smoking, we are not informed in NP which action the agent would prefer to perform. One-boxing therefore provides positive evidential relevance for getting $1million and static screening cannot be used in NP.

8) Problems with this theory

I believe that there are several problems with Horwich's argument that mean that it cannot fully explain OBI. In the Smoker's Fantasy, as we have seen, the desire for

34 Horwich (1985) pp433.
35 Ibid.
36 Ibid.
37 Ibid.
38 ibid

smoking is as strong evidence that the agent will get cancer as his actually smoking, and thus we can conclude that he has the smoking common cause. In general, in the other common cause cases, the desire to perform the dominant action (e.g. becoming an academic, not firing Smith) provides evidence that the agent has the common cause that correlates with that action. On the other hand, Horwich claims that, 'In Newcomb's Problem the causal chain leading to a given act is not mediated by any particular, easily recognizable, introspectible state. No characteristic desires or beliefs lead to one or another of the acts.'[39]

My first problem with Horwich's explanation for OBI (outlined in subsections 6 and 7) is why this should be the case in NP. The desire for smoking is the direct equivalent of the desire for $1000 in NP. If wanting to smoke makes it likely that the agent will get cancer, then wanting $1000 (a very easily recognizable introspectible state) makes it likely that there is no money in box b2. Desiring $1000 should provide the same evidence for having the two-boxing common cause as desiring smoking has for have the smoking common-cause. Thus, it seems questionable why we would believe that we do not know enough about the agent's beliefs and desires in NP to screen off the agent's act from the common cause.

My second problem is why an EDTer using static screening should definitely arrive at the 'two-boxing' answer in the other common cause cases. Consider my following argument which suggests that the agent's desires do not make it obvious whether the common cause obtains or not in these other common cause cases, and thus that the static defence cannot be used :

'In these other common cause cases the desires do not always make it obvious whether the common cause obtains or not. To see this, consider the Smoker's Fantasy. Imagine that *everyone* has a desire for smoking, and the agent is aware of this. In this case, an individual having a desire for smoking cannot assume that he has the smoking common cause, because everybody has such a desire. Some people with such a desire will be led to two-box and some will be led to one-box. Thus, the agent cannot be aware

39 Ibid pp438.

of whether his beliefs and desires will, when rationally deliberated on, lead to one-boxing or two-boxing.'

I suggest that these two problems can be combined to create a solution. Despite the desire for $1000 that is isomorphic with the desire for smoking, it is not obvious in NP which common cause obtains. This is because NP is interpreted in the exact way I have just described that would prevent screening in the other common cause cases if they were interpreted in the same way. In NP, it is not intuitively plausible to assume that the agents who one-box and the agents who two-box have different desires. This is because we make the assumption that *everyone* desires $1000. Thus, the desire for $1000 does not provide evidence for one common cause over another, since people with both common causes will desire money. Since the agent cannot know which common cause his beliefs and desires are indicative of, he cannot screen.

We could avoid concluding that some agents with the desire for $1000 one-box and some two-box, and thus hope to re-instate screening, by assuming that everyone has the two-boxing common cause, and thus that everybody would be inclined to two-box. As well as this being an unrealistic suggestion[40], it would not in fact solve the problem. Understanding the structure of the problem that I introduced in sub-section 2 will help with this argument. If this were the case then 1% of the population, all of whom have the two-boxing common cause, would one-box (this is given in the conditions of the problem), and thus 100% of people one-boxing would have the two-boxing common cause. This is not consistent with the description of the problem. On the other hand, if we just assume that everyone acts rationally and everyone two-boxes, then 1% of these people (in fact my discussion in sub-section 2 illustrated that this figure could be as high as 49.95% for the structure of NP to remain) will have the one-boxing common cause, and 100% of these 1% (or of these 49.95%) will two-box, which is again inconsistent with the conditions of the problem. Thus, we must accept that some of the agents with the desire for $1000 have the one-boxing common cause, and that some of them have the two-boxing common cause.

40 Indeed in some descriptions of NP the agent is described as watching people ahead of him taking both one-box and two-boxes, and the Predictor's predictions being highly accurate in both cases.

The other common cause cases, on the other hand are interpreted so that the agent's desires do provide evidence for which common cause obtains. It is assumed that it is the presence or absence of the desire for smoking that leads to rational deliberation concluding with smoking or not-smoking. Similarly, it is assumed that the desire to not fire Smith, and the desire to become an academic, would not be the same if the common cause were different . Horwich discusses this assumption in passing, 'he may know that the operative distinction between the individuals was a difference in their desires. In that case, he might well be able to determine what sort of desires would have led to the performance of A, and having done this, he can see by introspection whether or not he has them.'[41] It is intuitively plausible in the other common cause cases to believe that the agent's desires lead to his different actions, and are therefore indicative of the different common causes.

Thus, I have illustrated that the reason that static screening cannot be used in NP is not fully explained in sub-sections 6) and 7). I have shown that there is an underlying reason that enables us to verify why the agent cannot know his common cause in NP as a result of his beliefs and desires. He cannot know his common cause simply from his beliefs and desires, because both common causes are associated with the same set of beliefs and desires. Therefore I have expanded on Horwich's reasoning and illustrated why he is correct in claiming that, in NP, 'no characteristic desires or beliefs lead to one or another of the acts, so no such state can be employed as an epistemological screen.'[42] Thus one-boxing will still provide evidential relevance for obtaining $1million, and the EDTer would one-box. I now turn to discuss how the common cause does act, if not via the agent's desire for money.

41 Horwich (1985) pp436.
42 Ibid pp438.

Section 3: The Static Defence (Newcomb's Problem)

9) How do the common causes act in NP?

I have shown that, whilst static screening can be used in the Smoker's Fantasy (and the other common cause cases) as a result of the desire for smoking, the equivalent desire for $1000 does not allow agents to use static screening in NP. This is because some agents who have the desire for $1000 one-box, and some of them two-box. The desire for $1000 is therefore not indicative of either common-cause. However, illustrating this does not necessarily show that EDTers cannot use screening in NP. The common cause must be acting via something in order to cause the agent's different actions. If the agent could become aware of what the common cause is acting via, there would be potential for the EDTer to use static screening in NP. I thus turn consider how the common causes could be influencing the agent's actions, if not via this desire for $1000, and why these different actions of the common causes do not in fact allow the EDTer to statically screen in NP. I consider two possible interpretations.

a) The first interpretation of how the common cause acts is that it acts via beliefs and desires other than the desire for $1000 so as to control the agent's choice to either one-box or two box. Although the agent is fully aware of these beliefs and desires[43], he does not know which action they will lead him to perform. This is because, as Horwich claims, 'there are no characteristic desires or beliefs that lead to the choice of one act over another', and these beliefs and desires in NP are not 'easily recognisable.' It is hardly obvious what kind of beliefs and desires would trigger the difference between one-boxing and two-boxing. Unlike in the Smoker's Fantasy, where we know that a desire for smoking will lead to smoking, and a desire to not smoke will lead to not-smoking, in NP the agent does not know how his decision is being mediated.

43 Eells, Sober (1986) pp226 show that we must make this assumption, 'it must be supposed that the agent knows what his beliefs and desires are. Otherwise the evidential principle could not even be applied.'

Furthermore, claims Horwich, screening can only work, 'for decisions involving the simplest deliberations.'[44] In NP he believes that agents who make either decision, 'will include many people who achieved that result by the most contorted of deliberations.'[45] This makes it even harder for the agent to know what decision he expects he will make in light of his beliefs and desires, and thus makes it even harder for him to be aware of which common cause he possesses. Essentially, the complexity of NP means that the agent is unable to know purely on the basis of his beliefs and desires which common cause he has, and thus which choice he will end up making. For this reason, one-boxing would still have positive evidential relevance for receiving $1 million, and the EDTer would one-box.[46]

b) My second interpretation of how the common cause might act considers what we could discover about the action of the common cause if all agents in fact possess the same beliefs and desires, including the desire for $1000. It is such an interpretation that prompts the extensive literature on NP discussing what is *the* rational solution to NP. Such authors would presumably not be satisfied by the suggestion in a) above, that agents in fact have other beliefs and desires not stated in the problem which make one-boxing rational for some of them, and two-boxing rational for others. If, unlike the suggestion in a), all agents have the same relevant beliefs and desires, yet they reach different conclusions, it seems that we must suppose that some of the agents are performing irrationally.

I now intend to illustrate why screening could not occur in such a case, in which some irrationality needs to be postulated. In doing this, I show that the use of static screening in NP is self-defeating. If EDTers can use screening in NP, we would need to make the assumption that people who one-box are irrational. If we need to make the assumption that people who one-box are irrational, EDTers cannot screen in NP.

44 Horwich (1985) pp438.
45 Ibid.
46 Such a case is picked up again in sub-section 17, where I discuss Eells's further arguments for how screening could potentially occur in such a case.

The illustration of why screening cannot occur in such a case takes up the remainder of this section (sub-sections 9 to 14), since it is far from obvious why screening cannot occur. The proposed interpretation of NP is that the common cause acts via the agent's *irrationality* or *rationality* to yield the one-boxing or two-boxing actions. Thus, to show that screening cannot occur in such cases, it needs to be shown that the agent cannot be aware of his own rationality. If the agent knew he could act rationally, then he would know which common cause he had, and he would be able to screen.

10) Irrationality in NP

I now begin my description of how, if static screening could be used by the EDTer in NP, then NP would become a case in which static screening could not in fact be used. If static screening could be used by the EDTer in NP, and if the agent has no relevant beliefs and desires other than those stipulated in the problem, it would show that two-boxing would be the rational choice. This is because the use of screening would remove any evidential relevance that the acts could have for which state obtains. Every agent should then two-box, since every agent desires $1000.

The interpretation of NP that I am now discussing is therefore as follows:

1) 99% of people who have the two-boxing common cause two-box.

2) 99% of people who have the one-boxing common cause one-box.

3) 99% of people who have the one-boxing common cause have not performed rationally.

There are two ways in which the common cause could cause this irrationality.[47] One of these ways is that the common cause could affect the agent's deliberation (preventing him from screening), and the other is that the common cause could affect the action that the agent performs (preventing him from two-boxing although he intends to).

47 I am not considering the case where the common cause causes irrational beliefs. This would constitute the kind of case discussed at the beginning of the sub-section 9 ,where the beliefs (though irrational) and desires can lead under rational deliberation to the one-boxing and two-boxing common causes.

Eells refers to such outcomes, that are not in line with rational deliberation, as 'slips.'[48] Type 1 slips occur when an irrational intention is formed as a result of mistaken deliberation, whilst Type 2 slips occurs when the unintended act is performed.[49] What the statements above illustrate is that, if screening were able to function in NP, there would be a high correlation between the one-boxing common cause and the occurrence of slips. Thus, either something has gone wrong in these people's deliberations (Type 1 slips), or they have deliberated correctly and not been able to carry out the action they attempted to perform (Type 2 slips).

It certainly seems far too large a correlation to suppose that, of all the people reaching the irrational conclusion, 99% of them coincidentally have a certain common cause. This, it seems that we must assume:

EITHER that the one-boxing common cause causes the irrational conclusion to be reached, affecting the agent's deliberation or the agent's action directly; OR

that the one-boxing common cause is indicative of something that causes the irrational conclusion to be reached.[50]

There is a third potential causal structure, which is that the irrational action causes the one-boxing common cause, but this has been ruled out by the conditions of the problem. The first of the above two is the most natural to suppose,[51]but accepting either of these two claims would mean that one-boxers have an increased likelihood of performing Type 1 or Type 2 slips.[52]

Thus, given that only 1% of one-boxers are capable of performing the rational action, we must assume that the one-boxing common cause acts to cause some irrationality in the form of either Type 1 or Type 2 slips. I have described how the

48 Eells and Sober (1986) pp238.
49 Ibid.
50 As Eells (1989) pp12 claims, 'when a rational agent's subjective probabilities strongly correlate two items, then we can expect the agent to have causal beliefs... that would explain the relevant correlation.'
51 Ibid.
52 This of course is not to claim that two-boxers do not also slip, and that one-boxers who have slipped do not 'slip back.' These random and normal slips make up part of the 1% of people who's actions are not consistent with their common cause.

common cause could affect the agent's deliberations or actions. I now turn to formalise precisely why these considerations prevent an EDTer from screening in NP.

11) Static screening cannot be used:*when the assumption of Premise 4) is rejected, EDT can recommend one-boxing. 'Evidential decision theory can prescribe the wrong act.'*[53]

Eells states formally in Premise 4 (hereafter P4) of his argument that screening cannot occur unless the common cause acts only via the agent's beliefs and desires (Premise 4: 'The common cause does not affect his deliberation'). In the interpretation of NP that I am considering, this is false. I now describe why Eells believes that rejecting P4 would prevent screening from taking place.

That the common cause does not influence the agent's action (Type 2 slips) allows screening-off of the agent's common cause from his *choices* to be carried as far as screening off the agent's common cause from his *actions*. This is because, if the agent's *choices* have no evidential relevance to the common cause, 'the only way there could still be a significant positive correlation between A and O so that evidential decision theory will still prescribe act -A'[54] is if this equality is false:

Pr (A & choose-A & R & CC) = Pr (-A & choose A & R &CC)[55]

This equality states that the probability of Type 2 slips is entirely symmetrical. When the same common cause and the same beliefs and desires are present, the probability of slipping in either direction upon choosing an act is the same. However, in allowing the common cause to affect the agent's action, this condition no longer holds; the probability of one-boxing when you have chosen to two-box and you have the one boxing common cause is far greater than the probability of two-boxing when you have chosen to one-box and you have the one-boxing common cause. Thus, one-boxing will still provide positive evidential relevance for having the one-boxing common cause, and

53 Eells, Sober (1986) pp240
54 Ibid pp239.
55 Armendt (1988) pp328, equality playing the same role given in Eells, Sober (1986) claim 2) pp 234.

EDT will yield the one-boxing solution.

That the common cause does not influence the agent's deliberation (Type 1 slips) allows the screening off of the agent's common cause from his *choices*. If the common cause did affect the agent's deliberation we would be given reason to question the equality:

Pr (CC | choose A & R) = Pr (CC | choose -A & R)[56]

This equality claims that the probability of a certain common cause obtaining given that the agents holds certain beliefs and desires, is independent of the agent's choice. For example, this holds in the case where, given the desire for smoking, the probability of having the non-smoking common cause is the same whether the agent chooses to smoke or chooses not to smoke.[57]

If the one-boxing common cause influences the agent's decision-making process in the way described (causing Type 1 slips), then this inequality will not hold, since choosing one-boxing when holding the same desires as someone who chooses two-boxing is evidence for the one-boxing common cause. The probability of choosing one-boxing given the agent's beliefs and desires and the one-boxing common cause is greater than the probability of choosing one-boxing given the same beliefs and desires and the two-boxing common cause, since the two-boxing common cause will allow the agent to rationally evaluate his beliefs and desires and perform the action that follows from this evaluation, whereas the one-boxing common cause will not. Choosing to one-box would therefore provide the agent with positive evidential relevance for having the one-boxing common cause. EDT would recommend one-boxing. Thus, I have illustrated precisely why Eells believes that rejecting P4 would prevent screening.

Essentially the reasoning for why screening cannot occur when P4 is rejected is as follows: if something other than the agent's beliefs or desires affects what action the agent is going to perform, it is not possible for the agent to know solely on the basis of his beliefs and desires what common cause he has, which is required to use the static defence. Thus, the action that the agent performs will provide positive evidential

56 Armendt (1988) pp327.

57 This is the equivalent of Premise 10 in Eells's argument for screening, except for the agent's action has been replaced by the agent's choice.

relevance for having the common cause that correlates with it.

12) Static screening can be used: Argument 1)

I now turn to consider two different arguments that claim that Eells was mistaken in believing that screening cannot occur in cases in which P4 is rejected.

Although the common cause is acting via something other than the agent's beliefs and desires, as long as the agent could recognise this other effect of the common cause, he would potentially be able to screen. Premise 1) of Eells's description of static screening claims that the agent is Bayesian rational, and knows that he is Bayesian rational. Given this, it seems strange that Eells believes that screening cannot occur in a case in which one common cause causes irrationality. My first argument for why screening can still occur is thus that, if the agent is rational and knows this, he would also know that he is not an agent whose deliberations or actions could be influenced by the common cause. Thus, since he knows that he can act and deliberate rationally, he would know that he had the two-boxing common cause. The evidential relevance of one-boxing could then be screened off:

P (TB | R & TBCC) = P (TB | R & OBCC) where TB= two-boxing, TBCC= having the two-boxing common cause, OBCC = having the one-boxing common cause, R = the agent's beliefs and desires, and *knowledge of whether he is able to deliberate to and carry out the rational act.*

Thus, the agent's beliefs and desires combine with his knowledge of his own rationality to screen off the action from the common cause.[58] EDTers who know themselves to be capable of deliberating and acting rationally realise that one-boxing would not provide them with evidence that they have the one-boxing common cause, since, as a result of their rationality, they already know they have the two-boxing common cause. Thus, a rational EDTer would two-box.

It becomes clear that this argument does not work once we consider the

58 The agents knowledge of his rationality is not merely a subjective belief, and thus cannot just be considered another of the agents beliefs.

definition of rationality. An agent who is caused to perform Type 2 slips would not in fact be labelled Bayesian irrational.[59] Bayesianism is not concerned with whether an agent's intentions are carried out, 'there are many forms of Bayesianism, but the core idea is that a Bayesian agent is an agent whose beliefs and desires (or preferences) satisfy various formal coherence conditions. The most popular view also assumes that Bayesian agents update their beliefs by conditionalizing upon incoming evidence in order to maximise their expected utility.' For a Bayesian rational agent, 'knowledge and goals are subjective notions, constrained only by self-coherence and coherence with resulting intentions as constant updating of subjective probabilities occurs.'[60] The failure to carry out formed intentions could, for example, result from a lack of self-control, which is not the same as a lack of rationality. Thus there is no reason to suppose that, even if an agent could know that he is Bayesian rational, he should know whether he will perform Type 2 slips.[61]

On the other hand, an agent who Type 1 slips *is* deliberating irrationally.[62] Is it reasonable to suppose that a rational agent could know that he will not slip in this way? The one-boxing common cause acts only under the circumstances of NP (which the agent may have never come across before) to cause a mistake in the agent's reasoning or action. Whether this common cause is present is hardly something we can expect the agent to know. In fact, a rational agent certainly cannot know whether he will be caused to Type 1 slip, since even rational agents must allow for the possibility of themselves performing random Type 1 slips and thus deliberating irrationally[63]. If a rational agent must accept that he may occasionally perform a Type 1 slip without knowing this in advance, he could hardly be expected to know whether he will be caused by a common-

59 This raises a problem with Eells's argument for static screening (sub-section 5). He moves from the claim that the agent must be Bayesian rational (Premise 1) to the claim that two-boxing is the rational action (Conclusion). This implies that everyone who one-boxes is Bayesian irrational, which simply isn't true. Eells could only claim that those who choose to one-box are Bayesian irrational.

60 Pers.comm. David McCarthy.

61 Since these slips *always* occur when an agent with the one-boxing common cause chooses to two-box, we could question whether the agent actually has an intention to two-box. If the agent is not interpretable as having the intention to two-box, then he has formed an irrational intention and could be labeled Bayesian irrational.

62 There may be a temptation to consider that someone who is rational, but who has been 'caused' to act irrationally, is actually rational. However, such reasoning is mistaken. The gene in these common cause cases leads to an irrational intention *always* being formed in this situation. This is not a regular slip, it is in fact genetic irrationality. Although I use the word 'slip' for both mistakes, they are in fact very different.

63 Eells, Sober. (1986) pp236, 'slipping *is* a realistic possibility that a rational, human agent *should* acknowledge.'

cause to perform a Type 1 slip on a particular occasion. Thus, I reject the claim that a rational agent can know he will perform rationally. The rational agent must accept that he will randomly perform Type 1 slips without knowing this in advance, and thus he cannot know, when facing a certain decision, that he will deliberate rationally. One-boxing will therefore still have evidential relevance, since the agent does not know whether he will be caused to deliberate or perform irrationally or not.

I have shown that it is a mistake to suppose that an agent can know if he is going to be caused to perform irrationally or not. Yet, since Eells believes that an agent *can* know if he is rational (premise 1 of Eells's static screening argument in sub-section 5), why does he still maintain the belief that an agent cannot screen? I suggested that if an agent knows he is rational, he could know whether had the common cause that was causing him to deliberate irrationally. Eells, on the other hand, believes that if an agent knows he is rational, then he knows that there is *no possibility* of being affected. Eells thus believes that the rational agent knows not only that he can act rationally, but that if he had a different common cause he would also be able to act rationally. For this reason, Eells does not in fact believe that cases in which one of the common causes generates irrationality are appropriate cases for rational deliberation. I come to consider why I believe this is mistaken in sub-section 14.

13) Static screening can be used: Argument 2)

I now argue that Eells's argument for why screening cannot occur when P4 is rejected fails to acknowledge that not all evidential relevance can be gained by the agent's *active* decisions.

Consider the case where Type 2 slips are caused (i.e. where the agent is not able to perform the action he intended to perform) and, for simplicity, where there are no Type 1 slips. As I have shown, Type 2 slips prevent the screening-off of the common cause from the agent's choices being translated into screening-off of the agent's common cause from his actions (sub-section 11). Thus, if the structure of NP could be explained only by Type 2 slips, and not by Type 1 slips, we would at least be able to screen as far

as the agent's choices.

Eells claimed that in such cases the EDTer would still one-box, since the agent's action would still have evidential relevance. The one-boxing common cause causes the agent to one-box irrespective of his beliefs and desires, thus one-boxing will provide evidence that the agent has the one-boxing common cause. However, I now argue that since the agent cannot *actively* choose his *action,* but only his *choice*, screening as far as the agent's choice should yield the correct recommendations. Evidential relevance of an action cannot be gained by any decision that the agent makes.

Now, it is certainly true that the act of one-boxing will provide evidence to which the agent did not already have access via introspection.. If the agent, having deliberated rationally and formed an intention to two-box, then slips and one-boxes, this provides evidence that he has the one-boxing common cause.[64] However, no *active* choice that the agent makes can provide this evidential relevance. The agent can only choose his intention, and must then simply see whether he slips and one-boxes, which would then provide him with evidential relevance. When the agent chooses which action to try to perform it is true that the agent cannot be totally sure of which common cause he has. Yet, despite this, he has all the evidential relevance that it is possible for him to gain *actively*. Therefore, if the agent knew what choice his beliefs and desires would lead to when he deliberated rationally, knowing that Type 2 slips might occur would not affect his ability to screen as far as his choices. Though the agent's *actions* are not probabilistically independent of which common cause obtains, this lack of screening should not affect his choice of action, which would be independent of the states that obtain. Thus I disagree with Eells and Sober that, 'if the agent believes this [that type 2 slipping occurs] then... evidential decision theory can give the wrong answer.'[65]

Type 1 slips do still pose a problem to screening. The agent still cannot know what his beliefs and desires would lead to when he deliberated in the way that he considered to be rational, due to the possibility of these Type 1 slips. One-boxing would therefore provide positive evidential relevance that the agent was irrational,and thus that

64 If a slip occurs from the decision to two-box to the action of one-boxing the presence of Type 2 slips leads to the conclusion that this slip was in fact caused by the one-boxing common cause, rather than being a regular slip (i.e. just the agent with the two-boxing common cause making a mistake.

65 Eells, Sober (1986) pp239.

the agent had the one-boxing common cause so that the agent in NP with Type 1 slips would still one-box.

If the common cause did cause only Type 2 slips, then it would be possible for the EDTer to screen in NP as far as the agent's choices, thus recommending the choice of two-boxing. Yet, as long as the agent has to consider the possibility of the common cause affecting his deliberation, static screening cannot be used. I have thus shown that static screening cannot occur in NP. I showed in sub-sections 6, 7 and 8 that this was intuitively obvious to agents. In this section, I have now confirmed this intuition with technical details illustrating that the EDTer cannot statically screen in NP.

14) An appropriate case for rational deliberation? *'the agent, 'should not consider his decision problem to be one that is appropriate for the successful application of principles of rational decision.'*

Eells did not claim that the agent must believe that the common cause only influences his beliefs and desires (P4) in order to allow static screening to be used. Eells offers further arguments, purporting to show why cases in which P4 does not hold are not appropriate for rational deliberation.[66] If these cases are not appropriate for rational deliberation, my discussion of whether screening could be used by the EDTer if NP were such a case would be rendered pointless. I now illustrate why I do not believe that these further arguments are sound, and thus why the rejection of the assumption behind Premise 4 will not prevent NP being an appropriate case for rational deliberation.

First, in cases where Type 1 slips occur and the common cause affects the agent's deliberation, Eells claims that, 'we may suppose that it [the common cause] will not affect his deliberation, and that the agent believes this'[67] and that, 'this is plausible on the assumption that the decision maker is rational, and will make the rational decision whether or not he has the bad genetic condition - and that he believes this too.[68]' Essentially, Eells believes that if the common cause affected the agent's deliberation then

66 He thus believes that, in any case appropriate for rational deliberation, screening can occur.
67 Eells (1984) pp74.
68 Ibid.

the agent could not consider himself to be rational, and thus the situation would not be suitable for rational deliberation.[69] This is the underlying structure of Eells's argument:

1) For a case to be appropriate for rational deliberation the agent must be rational.

2) For a case to be appropriate for rational deliberation the agent must know that he is rational.

3) Therefore, for a case to be appropriate for rational deliberation it needs to be the case that the agent will make the rational decision whether or not he has the bad genetic condition, and that the agent knows this (from Premises 1 and 2).

Conclusion: For a case to be appropriate for rational deliberation, neither common cause can prevent the agent from deliberating rationally.[70]

This argument relies on this unstated premise, required to move from Premises 1 and 2 to Premise 3:

4) Being rational means the agent being able to know that he will make a rational decision whether he has the bad genetic condition or not.

Are these premises of this argument sound? Prescriptions of decision theory are meant to apply only to rational agents, thus Premise 1 is sound. We are not looking to offer prescriptions for irrational agents.

I now turn to Premise 4 of this argument. Eells thinks that, 'the assumption that an agent is rational is enough to ensure that the presence or absence of a common cause

69 Eells is not claiming that it is not possible for cases to occur where the action or deliberation are affected by the common cause, but is claiming that such cases would not be cases where it would be suitable for us to discuss prescriptions for rational deliberation.

70 The common cause could affect his deliberation and still leave the case appropriate for rational deliberation, if the affect just caused a different deliberation that also yielded the rational answer.

will not affect which decision rule is used',[71] This Premise is unsound. Being rational does not require that you are able to deliberate rationally if you have a different common cause to the one that you actually have. This would require a rational person to be able to counter-factually affirm that, with a different gene or brain structure, he would have been able to perform rationally. However, if a rational agent had had a different gene or brain structure he may well have been irrational; after all, it is likely that irrationality has genetic components. Therefore, rejecting P4 does not render the agent irrational as Eells suggests. An agent can certainly be rational without, 'being able to know that he will make a rational decision whether he has the bad genetic condition or not.'

I have already rejected Premise 2 in sub-section 12 above. Part of the reason for this rejection was that I did not believe that a rational agent would be able to know whether his deliberation would be caused to be irrational if he had had a different common cause. I considered it untenable that an agent could know this. However, I have now shown that this is not a condition of rationality, and thus that, for an agent to know he is rational, he need not know this at all. However, Premise 2 must still be rejected; as discussed in sub-section 12, it would still require the agent to know whether he will accidentally slip and deliberate wrongly. Since this is a possibility that all rational agents must accept, no rational agent can be sure that he will deliberate rationally.

Even though a rational agent cannot be sure that he will deliberate rationally, this does not mean that a rational agent finds himself to be in a situation which he considers inappropriate for rational deliberation. As long as there are in actual fact agents who can deliberate rationally, and as long as agents believe that they may be able to deliberate rationally, then deliberation is not pointless.

Turning now to turn to Type 2 slips, Eells claims that it is pointless for the agent to deliberate in cases where Type 2 slips occur, asking, 'why would you even bother deliberating between A and -A if you believed that, regardless of the deliverance of your deliberation, your act will be caused to agree with the outcome[72] that does obtain?'[73]

71 Eells (1981) pp320.
72 Common cause cases are described as having the cause, action and outcome (i.e. getting cancer, getting the job, there being $1000 in box b1 etc)
73 Eells, Sober (1986) pp240.

Thus, claims Eells, this would not be a case where rational deliberation is appropriate, and the agent should realise this, 'he should not consider his decision problem to be one that is appropriate for the successful application of principles of rational decision.'[74]

Eells is correct that, if both one-boxing and two-boxing were simply caused to correlate with the common causes irrespective of the agent's deliberation, then there would be little reason to deliberate. Yet, for these Type 2 slips, it is important to note that we need not believe that *both* actions are made to correlate with the appropriate outcome. Whilst we may need to assume that something is causing the one-boxers to be unable to carry out their intended action (if Type 2 slips rather than Type 1 slips are responsible for the correlation between the common cause and the performance of the irrational action), this assumption need not apply to two-boxers. Two-boxers can make choices in accordance with the principles of rational decision theory. As long as we can assume that a high proportion of the population are able to deliberate rationally[75], then those with the two-boxing common cause will choose to two-box, and in these cases no slipping will be caused by the common cause. Thus, there are still agents for whom deliberation is not pointless. As Armendt notes, 'asymmetric fallibility can fall far short of yielding inevitable action no matter what the course of deliberation.'[76]

Thus, whether the one-boxing common cause acts via Type 1 or Type 2 slips, there are agents who are capable of acting rationally, and thus for whom deliberation is not pointless.

15) Summary

I have considered Horwich's idea that EDT is responsible for OBI. Horwich claimed that screening could occur in the other common cause cases because it is obvious in such cases which common cause underlies the agent's beliefs and desires,

74 Ibid pp239-40.

75 Without this assumption we may have to suppose that the two-boxing common cause also has an influence over either the agents' deliberation or the agents' action in order to keep these agents acting in accordance with their common cause. If this were the case, then it seems that deliberation could be pointless. This is because, however the agent deliberated, either his deliberation or action would be caused to coincide with his common cause.

76 Armendt (1988) pp328.

whereas in NP the agent does not know which common cause underlies his beliefs and desires (sub-sections 6 and 7). I illustrated that the claim that the agent does not know which common cause underlies his beliefs and desires in NP was problematic, since the desires in NP give us as much information about the common cause as in all the other common cause cases (subsection 8).

I then explained why this need not be a problem. I illustrated that screening relies on a specific interpretation of the structure of the other common cause cases, an assumption that Horwich fails to acknowledge fully. NP, on the other hand, could never be interpreted with this structure, since the desire for $1000 is ubiquitous. Since the desire for $1000 is not indicative of one specific common cause, screening cannot occur (discussed in sub-section 8). Thus, despite NP and the other common cause cases having the same structure, this key 'ubiquitous desire' in NP means that the agent can become informed about his common cause from his beliefs and desires in the other common cause cases, but not in NP.

I then discussed two possible ways in which the common causes in NP could be acting, since they are not acting via the desire for $1000. It could be that the common causes are acting via other beliefs and desires, and that screening could fail to occur because the agent is not aware of which common causes these beliefs and desires stem from. Alternatively, it could be claimed that the one-boxing common cause in NP must cause either Type 1 or Type 2 slips (discussed in sub-section 9). I then explained in detail why screening would be unable to occur in any cases that involve Type 1 and Type 2 slips being caused by one of the common causes (discussed in sub-section 11). I developed two of my own arguments with a view to demonstrating that screening could still occur, but I illustrated why both arguments eventually failed (discussed in sub-sections 12 and 13). Thus, screening cannot occur in NP.

The last remaining problem for cases in which the common-cause causes Type 1 and Type 2 slips was that Eells claims that such cases do not constitute cases that are appropriate for rational deliberation. I refuted this claim in sub-section 14. Thus, I have expanded upon Horwich's explanation, and illustrated that as a result of using static screening, agents two-box in the other common cause cases, but one-box in NP.

16) Empirical evidence

I have illustrated, from a philosophical perspective, how it is understandable in NP that EDTers one-box. Yet there are various questions that could be posed that would help to determine the truth of the hypothesis that EDT explains OBI. For example, it would serve to find out whether those who support EDT do in fact believe, in the other common cause cases, that performing an action for evidential relevance is not possible since they already believe themselves to know which common cause they have.

Furthermore, other cases could be developed that could help to validate this hypothesis. For example, consider a case with the same structure as the Smoker's Fantasy, where the agent is deciding between smoking and not smoking. However, in this case the agent is not described as having a preference for smoking, but instead is told that if he smokes he gets $1000. According to my theory, this should, like NP, be more likely to yield a one boxing, 'not-smoking' answer than in the actual Smoker's Fantasy, since the desire for $1000 would not lead the agent to believe that he knew which common cause was present, unlike in the true Smoker's Fantasy where the desire for smoking leads the agent to believe that the smoking common cause is present.

I have thus offered an explanation for OBI. I now turn to consider further arguments that have been put forward to explain why screening can occur in NP. I aim to refute these arguments in order to affirm the explanation of OBI that I have derived above.

Section 4: The Dynamic Defence

17) Eells's dynamic defence

I have shown that EDTers cannot statically screen in NP. Despite this, Eells believes that screening can always occur and, in recognising that static screening leaves itself open to this kind of objection, he moves from his static defence to a more dynamic defence, 'Eells is moved by Horwich's Objection to abandon his initial static defence of evidential decision theory.'[77] Appealing to the dynamic defence purports to show that the agent can discover what conclusion his beliefs and desires lead to during the course of his deliberation, 'at some point before the moment of decision and from that point on he knows what his final credences and preferences will be and 'knows' what decision he will make of them.'[78] Thus, even if at the beginning of deliberation the agent does not know whether his beliefs or desires are of the type caused by the common cause or not, 'the required independence...should in fact, eventually...hold.'[79] Appealing to the dynamic defence therefore suggests that screening can occur in NP, and thus that the EDTer in NP should two-box. This would meant that EDT could not explain OBI.

The best way to illustrate how the dynamic defence should function for the EDTer is by example.[80] As we have seen, Eells does not believe that cases in which the common cause affects anything other than the agent's beliefs and desires are appropriate for rational deliberation (discussed in sub-section 14). He therefore believes that the dynamic defence is required for cases in which the common cause only influences the agent's beliefs and desires, but that these beliefs and desires are complex enough that the agent cannot tell purely from them which common cause he possesses ('the causal relationships between a person's antecedent physiological and psychological states and

77 Sobel (1991) pp151
78 Ibid pp156 from Eells (1984) pp88
79 Ibid pp149 from Eells (1984) pp78-79.
80 Full description taken from Eells (1984) pp83-92.

his subsequent action are often extraordinary complex, subtle, and difficult to recognise.'[81]) I discussed at the beginning of subsection how NP could be such a case.

For simplicity we assume that, during the course of deliberation, the agent receives no new information from the outside, and thus that the only new information available to him will be whatever he learns from the progress of his deliberation. The agent is confronted with the decision problem; he is fully aware of his beliefs and desires, and is aware that the only way in which the common cause can cause the symptomatic act is by influencing these beliefs and desires. Yet, imagine that he does not know whether these desires will lead him to act a1or a2, and thus at this point both actions would give him evidence about which common cause he has. In this case, 'a2 (a1) is highly positively (negatively) subjectively probabilistically relevant to C, the very bad causative factor... and a2 dominates a1 with respect to..the agent's preferences.'[82]

The agent begins by calculating the CEU of a1 and a2 Let us say that initial deliberation yields the result that a1 has a higher expected utility than a2 (this is what the EDTer's first calculation would yield). As a result of this calculation, the agent's subjective probability of him choosing a1 increases. This in turn provides him with evidence that, 'his beliefs and desires are of such a kind that rational deliberation in the light of them will, in the end, favour act a1.'[83]

The agent then recalculates. Under recalculation, the expected utility of a1 will be lower than before since, 'a1 has lost some of its negative evidential relevance to C [the very bad common-cause] and a2 has lost some of its positive evidential relevance to C.'[84] In other words, the action of a1 would no longer provide strong evidence to the agent that he does not have C, since he already believes this to be likely. As the recalculations continue, a1 may for a while continue to have a higher expected utility than a2. Yet, as these calculations continue, the agent notes that it becomes even more likely that he will choose a1, and even less likely that he has C. The agent has started to gain information about which common cause his beliefs and desires stem from.

81 Horwich (1985) pp436.
82 Ibid pp87.
83 Eells (1984) pp87.
84 Ibid.

Eventually, C, 'will be sufficiently independent from the acts'[85] that the expected utility of a2 will increase above that of a1. This is because the agent is almost certain that he does not have C, and thus that he will not get the bad outcome of C. He would then prefer to be performing a2, since a2 dominates a1. As the expected utility of a2 increases, the agent comes to believe that he will choose a2. He therefore believes that his beliefs and desires are of the type that rationally lead to a2, and thus that they are caused by C. As these calculations continue they increase the degree of independence of the common cause from the acts. Eventually, the agent believes that he will choose to a2 and has C, and thus that choosing a2 will not provide him with any evidential relevance for having C. Choosing a2 is not bad news, since the agent has already received the bad news. The agent should choose a2.

Thus, as Eells says, 'Horwich's objection is thus answered by supposing that the agent continually calculates CEU, with appropriate alterations in his subjective probabilities in light of the results of previous calculations.'[86] As a result of his deliberation, the agent gains the information that caused Horwich's objection, that is from which kind of common cause his beliefs and desires stem. Before he acts, the agent knows that he has C, since he has been led to a2 as a result of rational deliberation, and thus the actual act performed will not provide him with any evidential relevance. I now turn to consider whether the EDTer could use the dynamic defence in NP.

18) The dynamic defence and Newcomb's Problem.

Eells has answered Horwich's objection by showing that dynamic screening can be used in a case in which we may know our beliefs and desires but we do not know which common cause they stem from. Thus, the EDTer can use screening in NP, if it is interpreted as described in a) of sub-section 9.

However, I also attempted to show that screening would be unable to occur if NP were interpreted as a case in which the one-boxers were being caused to deliberate or act

85 Ibid.
86 Ibid pp92.

irrationally (sub-section 9). Eells did not set out to show that screening could occur in such a case, since he believes that cases in which the common cause does not act solely via beliefs and desires do not constitute cases that are suitable for rational deliberation (sub-section 14). Having illustrated this belief to be mistaken (discussed in sub-section 14), I now turn to consider whether the EDTer can use the dynamic defence in NP, despite the Type 1 slips.[87]

We are only concerned with rational agents in decision theory. Thus, consider the deliberations of the rational agent. He may recognise that some of the time his common cause will cause him to deliberate irrationally, and thus to be led to believe that an irrational choice of action is rational.[88] For the rational agent, deliberation will occur just as described in the dynamic defence, which will lead to the two-boxing conclusion. This causes the agent to believe that he has the two-boxing common cause. At this point, the agent may be aware that this deliberation could have been rational, or that he could have been caused to reach an irrational conclusion despite considering it to be rational.[89] Either way, he believes himself to have the two-boxing common cause (whether two-boxing is in fact the rational action or not), and thus choosing to one-box will not provide him with any evidence for him having the one-boxing common cause.

Thus, the process of deliberation allows an agent to discover which action seems rational for him to choose. Choosing to perform not the action that appears rational to the agent, but the action opposite to it can then never provide him with evidential relevance, because, if he had the common cause equivalent to the opposite action, choosing the opposite action would appear rational. Evidential relevance cannot be gained from performing an act that the agent sees as irrational.[90]

Thus, dynamic screening can occur in NP to yield the two-boxing answer. EDT therefore recommends two-boxing for both NP and other common cause cases. Thus far,

87 I have already illustrated how screening can still occur as far as the agent's choices despite Type 2 slips in sub-section 13, and thus do not consider them here.

88 Here I have made that assumption that, when the agent is caused to deliberate irrationally, he believes he is deliberating rationally. If, on the other hand, as he finds himself deliberating irrationally, he knows that he is deliberating irrationally, then such an irrational agent could simply be ignored.

89 As we have seen it seems reasonable to assume that everyone desires money, thus meaning that all agents could be aware of the caused irrationality of some agents.

90 These considerations do not change the thought process behind the rational agent dramatically. The thought process behind the rational agent's deliberation will remain similar to what we have already seen. He reaches his conclusion of two-boxing via what he believes to be rational deliberation, and he thus realises he has the two-boxing common cause.

our table of information regarding NP therefore looks like this:

	EDT + screening recommends	Do we have an intuition to one-box?	Can screening occur?
Newcomb's Problem	Two-boxing	Yes	Yes
Other common cause cases	'Two-boxing'	No	Yes

Figure 9: An illustration of why OBI is still unexplained.

As this table shows, once we consider dynamic screening, it seems that EDT cannot explain OBI. Whilst the claim that 'screening occurs in other common cause cases meaning that EDT recommends two-boxing in these cases' (claim a in sub-section 6) has been vindicated, the claim that 'screening cannot occur in NP (claim b in sub-section 6) is now problematic.

However, I now argue that the dynamic defence would not convince an EDTer in NP that he should two-box, since using the dynamic defence of screening presupposes the irrationality of EDT.

19) The circularity of the dynamic defence

I believe that the dynamic defence as described by Eells is invalid for the purpose for which it was developed. I argue that using the dynamic defence in fact presupposes the irrationality of EDT.

Eells uses the dynamic defence to show that, for EDTers, choosing to one-box

will not provide any evidential relevance over and above what they would already become aware of during the course of their deliberation. Thus, two-boxing is the rational act. However, in attempting to do this, Eells takes a stand on the kind of deliberation that he believes to be rational. He presupposes that *continual* conditional expected utility maximisation (CCEUM) is the rational method of deliberating[91] His argument therefore no longer applies to, 'some principle of expectation maximisation' (Premise 2 of the static defence), but is restricted to those who believe CCEUM to be the rational mode of deliberation. The EDTer does not embrace CCEUM. Therefore, the EDTer is excluded from using the dynamic defence. Using the dynamic defence to convince and EDTer that the choice he wants to make is irrational is circular, since it presupposes that it is irrational to EDT.

Eells's dynamic defence goes as follows:

1. You believe that one-boxing will give the highest pay-off, and thus you believe you will choose to one-box. You therefore believe you have a high probability of having the one-boxing common cause and of becoming rich.

2. Upon realising that you have the one-boxing common cause, you realise that you should choose to two-box in order to gain the extra $1000.

3. You have then been rationally led via CEU maximisation to two-boxing and therefore you must, since the common cause acts via rational deliberation, conclude that you are highly likely to have the two-boxing common cause.

4. Once you believe that you are a two-boxer, the pay-off from one-boxing is lower than that of two-boxing, and thus it is rational to choose to two-box.

The EDTer does not agree that this is the correct way of calculating the rational act. CCEUM in fact incorporates the reasoning seen behind the dominance argument of CDT. The dominance argument claims that you should two-box since, whether you have

91 Eells (1984) pp83-85 spends some time discussing the rationality behind its use.

the one-boxing common cause or the two-boxing common cause, you are better off two-boxing. EDT on the other hand claims that, since one-boxing provides favourable evidence that you are a one-boxer, the agent should one-box.

It may seem mistaken to claim that CCEUM uses dominance reasoning; since, rather than claiming that whatever common cause you have you should two-box, the argument actually *discovers* that you are a two-boxer. Even the EDTer agrees that if you are a two-boxer, you should two-box. However, CCEUM is only able to claim that you are a two-boxer, and thus that it is rational to two-box, as a result of dominance reasoning . CCEUM relies on dominance reasoning in comparing the pay-off of two-boxing vs one-boxing if you have the one-boxing common cause (stage 2 of Eells's dynamic defence outlined above), and the pay-off from two-boxing vs one-boxing if you have the two-boxing common cause (stage 4), in order to claim that two-boxing is the only rational choice. The EDTer, on the other hand, carries out his calculations by comparing the pay-off from one boxing given that you have the one-boxing common cause, with the pay-off from two-boxing given that you have the two-boxing common cause. It is not disputed in the reasoning of the dynamic defence that, if you find yourself choosing to one-box, there is a high probability that you have the one-boxing common cause, and if you find yourself choosing to two-box there is a high probability that you have the two-boxing common cause. The EDTer bases his calculations on these facts alone.

Thus, the use of CCEUM in the dynamic defence is essentially the use of dominance reasoning (or CDT), which, as we saw in sub-section 1, is at loggerheads with EDT.

Eells's dynamic defence contains within it, although not explicitly stated, two explanations of why an EDTer should not consider that choosing to one-box would provide him with positive evidential relevance for getting $1million. Considering these explanations helps to illustrate further the circularity of using the dynamic defence to address an EDTer:

1) If the one-boxer deliberates and concludes that he will one-box, then he already knows that he has the one-boxing common cause. Thus, actually choosing to one-box does not provide the agent with any positive evidential relevance for there being $1million in b2, since the agent already has this evidential relevance. Thus, claims the CCEUMer, once the EDTer concludes that he will one-box, and thus is a one-boxer, surely he must two-box?

If the EDTer found out for certain that he was a one-boxer and that there was $1million in box b2, admittedly he would agree that two-boxing was rational. Yet, what step 3 of Eells's argument shows is that, even if you are fairly certain you are a one-boxer, as soon as you conclude from this that it is rational to two-box, you must become fairly certain that you are a two-boxer.[92] Therefore, though the EDTer realises that there is a high chance that he has the one-boxing common cause, he will not then choose[93] to two-box. Choosing to two-box when the agent is fairly certain he is a one-boxer would yield an expected pay-off that he knows to be lower than his current pay-off; namely the pay-off of two-boxing whilst believing that he is a two-boxer. Eells is correct that one-boxing does not provide evidence for $1million being in box b2. It is rather the agent's deliberation i.e. his calculations of evidential expected utility, that provides this evidence[9495] However, the agent needs to actually make the choice of one-boxing to 'cash in' on the evidential relevance that he is already aware of. If he comes to believe that two-boxing is rational, he will lose the evidential relevance that he has the one-

92 Eells doesn't always distinguish between *deciding* to perform the other act, and *slipping* to perform the other act. For example in Eells (1989) he claims that once you've concluded that you will go for popcorn and thus that there is popcorn available, if you then didn't go for popcorn you would be missing out on the available popcorn. This is true if the agent randomly *slips* and performs the action that he did not intend to perform. However, if the agent *rationally decides* not to go for popcorn, this indicates that a different common cause is present and thus that there would not be any popcorn present.

93 Though if he slipped he would still be considered a one-boxer.

94 Eells (1989) pp15 emphasises this with regard to the popcorn problem, 'it is not the act G that gives the evidence that P, but rather it is the outcome of deliberation that provides this evidence', and 'the point is that in the process of deliberation, the agent acquires evidence to the effect that G would cause him to get popcorn; this evidential expected utility maximiser does not reason that G is good because this act would be good evidence for the availability of popcorn.'

95 Eells ibid claims that this means that , 'there is nothing that the agent has decided to do- nothing that the agent has control over- which he decided to do in order to provide himself with evidence that P: the evidence derives from calculations of evidential expected utility.' If EDTing is caused by the common cause as suggested earlier this would be true. However, once we realise that this need not be the case, (sub-section 21) surely we have some power over the kind of deliberation that we view as rational? The EDTer would presumably choose to EDT!

boxing common cause and gain positive evidential relevance for having the two-boxing common cause.

In fact, since the EDTer sees two-boxing as irrational, if he did then choose to two-box this would not provide any evidential relevance for having the two-boxing common cause. As I showed in sub-section 18, when something is viewed as irrational then it will not provide the agent with evidential relevance, because its irrationality already informs us that the agent does not have the common cause that leads towards choosing such an action. That he finds two-boxing irrational does indeed tell us that he has the one-boxing common cause. Yet, doing something that one sees as irrational would obviously be irrational.

Continual conditional expected maximisation suggests that the move away from one-boxing must be rational; one boxing is not stable since two-boxing, given the one-boxing common cause, gives a greater pay-off. Since CCEUM features continual deliberation relative to the most likely current state, it always leads to stable solutions.[96] EDT on the other hand makes a static comparison, and thus stability does not play a role in its prescriptions.[97] The EDTer sees the new decision to two-box and the evidential relevance that comes with it as simultaneous.

2) Once you have deliberated and decided that you should choose to two-box, one-boxing is irrational according to CCEUM because it yields a lower pay-off relative to the current state of having the two-boxing common cause.

What this illustrates is that screening only discourages from choosing to one-box those agents who believed that choosing to one-box was irrational anyway. The reason that choosing to one-box would not provide any evidential relevance once the agent has decided that he will choose to two-box, is simply that one-boxing seems irrational as a result of the CCEUMer's dominance reasoning. Since the agent views two-boxing as rational, one-boxing will provide no evidential relevance to the rational agent.

96 Sobel (1990) discusses the notion of stability in rational decision making. A decision is stable if, relative to what the agent believes at that time, there is an action that would yield a higher pay-off than the one he is taking.

97 This could explain why the dynamic defence was not noted by Horwich; since it relies on such a specific mode of reasoning, and a mode of reasoning that is in fact not available to the EDTer.

It is true that for those using dominance reasoning, who conclude that two-boxing is rational, one-boxing would have no positive evidential relevance for there being $1million in b2. Thus, the CCEUMer should indeed stick to two-boxing. Using the dynamic defence can show why an agent who already believes EDT to be irrational would be provided with no reason to one-box. However, as a result of his calculations (comparing the pay-offs of two-boxing, believing yourself to be two-boxer, and of one-boxing, believing yourself to be a one-boxer), the EDTer sees one-boxing as rational. As a result, choosing to one-box does, for the EDTer, fulfil the evidential relevance that he has already become aware of as a result of his deliberation.

Thus, the appeal to the dynamic defence fails to illustrate that both EDT and CDT should converge to the same answer of choosing to two-box. Rather, using the dynamic defence presupposes the irrationality of EDT. Thus, screening cannot occur in NP. Again, it seems that EDT can explain OBI as illustrated below:

	EDT recommends	Do we have an intuition to one-box?	Can screening occur?
Newcomb's Problem	One-boxing	Yes	No
Other common cause cases	'Two-boxing'	No	Yes

Figure 10: An illustration of how OBI has so far been explained by EDT.

Unfortunately, however, one final consideration in the next sub-section leads me to conclude that screening cannot in fact occur even in the other common cause cases.

20) The circularity of the static defence

Recognising that the dynamic defence is circular leads me to question how the static defence avoids this circularity. In fact, it does not. Using the static defence to convince the EDTer that a certain choice is rational is a circular argument, since it presupposes the irrationality of EDT. This defeats Horwich's argument that explains why screening can occur in the other common cause cases whilst not occurring in NP, since it means that the EDTer cannot employ the screening defence in any cases. Therefore, EDT cannot explain OBI.

Horwich uses the static defence (described by Eells) to illustrate that, in cases such as the Smoker's Fantasy, the EDTer is not committed to prescribing the seemingly irrational abstinence from smoking. The argument, as we have seen, is that the agent's desire for smoking indicates that he has the smoking common cause, 'there is a reasonable inference from the statistical data: namely, that there is a high correlation between cancer and having the inclination to smoke.'[98] Thus, as a result of this, both the CDTer and the EDTer would realise as a result of introspection that they have the smoking common cause. Given this, the rational choice of action is to smoke. The EDTer recognises that there is no longer any point in him abstaining, since he already knows that he will get cancer.

In fact, this argument in the preceding paragraph presupposes CDT. Therefore, the EDTer would not following the reasoning behind the argument. Consider these claims that we have already considered, and are made in using the reasoning process of static screening:

'preferring the academic life to the athlete's life should be as strong evidence for the tendency [to pursue an academic career] as choosing the academic life,'

'there is a reasonable inference from the statistical data: namely, that there is a high correlation between cancer and having the inclination to smoke.'

98 Horwich (1985) pg433.

The common cause is supposed to act via the agent's beliefs and desires such that, in light of rational deliberation, the agent will reach the relevant outcome. Thus the assumption here is that the way the smoking common cause acts is to make the agent desire smoking, so that he will smoke. The non-smoking common cause causes him to not desire smoking, so that he won't smoke.

However, only someone who believed CDT to be rational would conclude that an agent's desire for smoking will lead the agent to smoke. For the EDTer, the desire to smoke would not be indicative of the smoking common cause. As we have seen, whilst the CDTer attends to the *consequences* of actions in deciding what to do, the EDTer attends to both the consequences of the actions, and what his action would be *evidence* for. Therefore, to the EDTer, the fact that smoking would be evidence for dying of cancer which he strongly prefers not to do is, to him, as strong evidence that he won't smoke as actually not smoking. The sort of desire that would make him believe that he would smoke would be a non-aversion to dying from cancer.

Thus, the initial conclusion reached by using the static defence was mistaken. The EDTer would see the desire to avoid cancer as indicative of the non-smoking common cause, since rational deliberation in light of such a desire would lead him to not-smoke. Yet the use of the static defence was intended to show that, whichever common cause the agent had, if he knew his common cause before acting, the action would no longer provide evidential relevance; the agent did not necessarily have to discover that he had the smoking common cause. Indeed, Horwich seems to make this line of reasoning obvious, 'let him use his beliefs, his desires, and the evidential principle to determine which act would be rational.'[99] So surely, because the EDTer knows that he is not a smoker, he should then smoke because, 'belief in the non-causal, evidential relevance of his act will become extinguished'[100]? This returns us to exactly the same explanation for why the EDTer would not do this seen with my discussion of the dynamic defence, the only difference being that in this case the agent discovered what his common cause was before deliberation. As I showed, the EDTer would still not-smoke, in order to avoid the positive evidential relevance of any other decision-making

99 Ibid pp437.
100Ibid pp438.

process. As with dynamic screening, using static screening only discourages not-smoking for those who believed not-smoking to be irrational anyway. It assumes that smoking is rational in light of the agent's desires in order to show that two-boxing is in fact rational in light of those desires.

Thus, the mistake of reasoning that makes using the static defence to address an EDTer circular, is attending to fulfilling desires that could be fulfilled by the acts' consequences, and not attending to fulfilling desires that could be fulfilled by what the acts gave evidence for. Whilst the CDTer would think it was rational to smoke when smoking was desired, and not when smoking was not desired, the EDTer would think it was rational to not smoke when not dying of cancer was desired, and rational to smoke only if not dying of cancer was not desired. Since the desire to avoid cancer is much stronger than the desire to smoke, the desire to smoke or not will make no difference to the EDTers' conclusions, and thus will not allow the EDTer to infer anything about his common cause. The CDTer believes that the common cause is acting via his beliefs and preferences in the light of rational deliberation, which he believes to be CDT. Thus, if he desires smoking, it seems to him that the common cause is leading him to choose to smoke in light of rational deliberation. On the other hand, the EDTer also believes that the common cause is acting via beliefs and preferences in light of rational deliberation, which he believes to be EDT. Thus, a desire to not die from cancer leads him to believe that he will choose to smoke in light of rational deliberation.

The original form of EDT sometimes yielded irrational answers, and Eells attempted, in his descriptions of screening, to give an analysis showing that EDT is not committed to these irrational answers. However, both static and dynamic screening violate the tenets of EDT, and thus cannot serve the purpose for which they were developed. Therefore, whilst a lack of screening explains how the one-boxing intuition would persist for EDTers in NP, it does not explain why EDTers do not 'one-box' in the other cases. EDT yields answers that seem entirely irrational (discussed in sub-section 3) in the other common cause cases.

Thus, at the end of this thesis we are left with this position:

	EDT recommends	What is our intuition?	Can screening occur so that 'two-boxing' is illustrated to be rational?
Newcomb's Problem	One-boxing	To choose to one-box	No
Other common cause cases	'One-boxing'	To choose to two-box	No

Figure 11: An illustration of why OBI could not e explained by EDT.

EDT cannot explain OBI. Having reached such a negative conclusion, in Appendix D I offer a brief description of the kind of work that could be undertaken to explain OBI using CDT rather than EDT.

21) Consequences

It is worth re-considering my earlier discussion in the light of what I have now established. I have now illustrated that using the static defence presupposes CDT. Presupposing CDT means that smoking seemed rational in the light of the desire to smoke, when actually someone with EDT would not smoke in the light of the desire to smoke. In the same way, presupposing CDT meant that only two-boxing seemed rational in the light of the desire for $1000. This led to me postulate either that those with the one-boxing common cause had some extra desires and beliefs that led one-boxing to be rational, or that the one-boxing common cause was causing them to deliberate or act irrationally (sub-section 9).

However, we now have an explanation for why some people one-box in the light of the desire for $1000. It is not that they have extra unmentioned beliefs and desires, nor that they are being caused to deliberate or act irrationally, but rather that, for the EDTer, one-boxing is rational. The advocate of static screening believes that as soon as the EDTer recognises that he has the one-boxing common cause, he should conclude by two-boxing. For this reason, one-boxing was presumed to be irrational. However, those who think that it is rational to adopt EDT will one-box in NP, and will not be persuaded to two-box by knowing that they have the one boxing common-cause. It is true that the EDTer knows that he has the one-boxing common cause, but he could not continue believing this if he were to two-box. After all, if an EDTer who reasons in this way will choose to two-box, the Predictor will have predicted him to choose to two-box. The EDTer would therefore one-box. This reasoning is not an irrational deliberation that has been 'caused' by the common cause, though it may appear to be so to the CDTer (and thus to the advocate of static screening).

Those who think that it is rational to adopt CDT will two-box. The CDTer could use the static defence to further convince himself that he is correct to two-box, since one-boxing, which he sees as irrational, would, for him, provide no evidential relevance for having the one-boxing common cause.

Thus, the EDTer will one-box and the CDTer will two-box.[101] If I am correct in arguing that the static defence would never be used by the EDTer, since it presupposes CDT, it makes sense to suggest that the common cause in NP is something about the agent (for example a brain structure) that allows the Predictor to determine whether the agent is likely to use EDT or CDT..[102]

101 Although I do not have room to fully discuss it here, the conclusions I have reached unfortunately make the structure of the other common cause cases *more* complicated. There is a correlation between those with the smoking common cause and those who CDT, and those with the non-smoking common cause and those who EDT. It seems that from this it would need to be assumed that the common causes in such cases are influencing the agent's deliberation.

102 Another interesting conclusion of there being two different types of 'rational' agent is that, even if an EDTer sees that lots of people with a desire for smoking have ended up smoking, he need not conclude that he has the smoking common cause. The common cause affects the way he deliberates, not his desires, and thus to discover his common cause he could not simply observe what other agents with similar desires do (a suggestion that Horwich (1985) pp436 makes).

22) Conclusion

I have investigated Horwich's suggestion that OBI can be explained using EDT. I began by showing why the arguments for static screening could seemingly be followed by EDTers in the other common cause cases, whilst not in NP, expanding on Horwich's suggestions and yielding a seemingly positive conclusion to the suggestion that EDT explains OBI.

Despite this, Eells suggests that dynamic screening should yield the two-boxing solution for the EDTer in NP. However, I then showed that the arguments of the dynamic defence in fact presuppose the irrationality of EDT in order to illustrate that one-boxing has no evidential relevance. Thus, in NP the EDTer would still one-box. I then argued that the EDTer could not in fact use the static defence in the other common cause cases, since the arguments of the static defence also presuppose the irrationality of EDT. Thus, in the other common cause cases the EDTer would also 'one-box'. Horwich's argument cannot explain why agents 'two-box' in the other common cause cases and thus EDT cannot explain OBI.

Word count: 16,490 (including appendix, excluding tables).

Acknowledgments

Thank you to David McCarthy, who has supervised this thesis.

Bibliography

Armendt, B. (1988), "Impartiality and Causal Decision Theory", *Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1: 326-336.

Burgess, S. (2004), "The Newcomb problem: An unqualified resolution", *Synthese* 138: 261-287.

Eells, E. (1982), *Rational Decision and Causality.* Cambridge: Cambridge University Press.

Eells, E. (1984), "Metatickles and the Dynamics of Deliberation" *Theory and Decision* 17(1): 71-95.

Eells E, Sober E (1986) , "Common Causes and Decision Theory" *Philosophy of Science* 53 (2): 223-245.

Eells, E. (1981), "Causality, Utility and Decision" *Synthese* 48: 295-329.

Eells, E. (1989), "The Popcorn Problem: Sobel on Evidential Decision Theory and Deliberation-Probability Dynamics" *Synthese* 81 (1): 9-20.

Horwich, P. (1985), "Decision Theory in Light of Newcomb's Problem", *Philosophy of Science* 52 (3):431-450.

Hurley, S.L (1991), "Newcomb's Problem, Prisoner's Dilemma, and Collective Action", *Synythese* 86 (2).

Joyce, J. (1999), *The Foundations of Causal Decision Theory* New York: Cambridge University

Press.

Levi, I. (1975), " Newcomb's Many Problems", *Theory and Decision* 6: 161-175.

Lewis, D. (1981), "Causal Decision Theory", *Australasian Journal of Philosophy* 59 (1): 5-30.

Ninan, D. (2006), "Illusions of Influence in Newcomb's Problem" presented at the 7th Annual Princeton-Rutgers Graduate Student Philosophy Conference.

Nozick, Robert (1969), "Newcomb's Problem and Two principles of Choice," in N. Rescher (ed), *Essays in Honor of Carl G. Hempel.* Dordrecht: Reidel, 114-47.

Nozick, Robert (1993), *The Nature of Rationality.* Princeton: Princeton University Press.

Nozick, Robert (1997), *Socratic Puzzles.* Harvard University Press.

Skyrms. Brian (1980), *Causal Necessity* Yale University Press.

Slezak, P (2005), "Newcomb's Problem as Cognitive Illusion" Proceedings of the 27th Annual Conference of the Cognitive Science Society, Bruno G. Bara, Lawrence Barsalou & Monica Bucciarelli eds., Mahway, N.J.: Lawrence Erlbaum, pp. 2027-2033

Sober E, Eells E (1986), "Common Causes and Decision Theory," *Philosophy of Science* 53 (2): 223-245.

Sobel, JH (1991), "Non-dominance, third person and non-action Newcomb Problems, and Metatickles", *Synthese* 86 (2): 143-172.

Sobel, JH (1990), "Maximisation, Stability of Decision, and Actions in Accordance with Reason", *Philosophy of Science* 57(1): 60-77.

Appendix A

P (Prediction two-boxing | Agent two-boxes) = 99% and

P (Prediction one-boxing | Agent two-boxes) = 99% which can be illustrated as:

	Predictor predicts one-box	Predictor predicts two-box
Agent one-boxes	0.99T	
Agent two-boxes	0.01T	0.99T

Where T= total probability of two-boxing.

From this, we wish to infer that P(Agent two-boxes | Prediction of two-boxing) = 99%.

With this additional information the situation can be illustrated as:

	Predicts one-box	Predicts two-box	Total
One-boxes	0.99T	0.01T	
Two-boxes	0.01T	0.99T	T

Since total probability of two-boxing + total probability of one-boxing = 1

Total probability of one-boxing = 1 – probability of two-boxing = 1-T.

Thus the final table appears like this:

	Predicts one-box	Predicts two-box	Total
One-boxes	0.99T	0.01T	1-T
Two-boxes	0.01T	0.99T	T

Taking the one-boxing row:

0.99T + 0.01T = 1-T

T = 1-T

T = 0.5.

Appendix B

If you one-box, your expected outcome according to Argument 1 = P (Predictor predicts that you will one-box | You one-box) (1million) + P (Predictor predicts that you will two-box | You one-box) (0)

If you two-box your expected outcome according to Argument 1 = P (Predictor predicts that you two-box | You two-box)(1000) + P (Predictor predicts that you will one-box | You two-box)(1001000) which can be re-written as:

P (Predictor predicts that you two-box | You two-box)(0) + P (Predictor predicts that you will one-box | You two-box)(1000000) +1000.

For Argument 1 to hold, the expected outcome for one-boxing needs to exceed that of two-boxing.

Setting the equations equal to each other we get:

P (one | one) (1million) = P(one | two) (1000000) + 1000

We thus need to solve the following simultaneous equations:

1million a = 1million b + 1000

a + b = 1

Solving the simultaneous equations:

b = 1-a

1million a = 1million (1-a) + 1000

1million a = 1million-1million a + 1000

2million a = 1001000

a = 1001000/ 2000000

a = 0.5005 QED.

Appendix C

I have shown in Appendix A that P (Predictor predicts that you one-box | you one-box) need only be greater than 0.5005 in order for EDT to recommend one-boxing. Thus, in this case:

P (Predictor predicts that you one-box | you one-box) = 0.5005

P (You one-box | Predictor predicts that you one-box) = 0.99

P (You two-box| Predictor predicts that you two-box) = 0.99

Thus the table making up the population is as follows, where T this time represents the total number of one-boxing predictions.

	Predicts one-box	Predicts two-box
One-boxes	0.990T	0.989T
Two-boxes	0.010T	97.814T

The percentages in the population are therefore:

	Predicts one-box	Predicts two-box
One-boxes	0.992%	0.990%
Two-boxes	0.010%	98.008%

Thus, these population percentages represent the most extreme percentages that can exist in the population so that Argument 1 can hold.

Appendix D

As we have seen, in its calculations EDT takes account of any probabilistic dependence between acts and states, whereas CDT only takes account of *causal* dependence between acts and states. This thesis has illustrated that EDT cannot be responsible for OBI. If EDT cannot recommend 'two-boxing' in the other common cause cases, this is certainly a blow to EDT, since 'one-boxing' seems obviously irrational. Another way of explaining OBI is to suggest that no-one intends to use EDT, since it is obviously irrational in recommending agents to act to provide evidence for something that they cannot have any influence over. Agents thus follow the prescriptions of CDT. The only reason that agents actually use what is *in fact* EDT in NP is that they make the mistake of believing that there are causal connections between their action and the pre-existing state, a belief that they do not hold regarding the other common cause cases.

Nozick highlights a difference between the cases that explains this mistake. He claims that the way agents use to tell if an action affects a state is if, 'the action is referred to in an explanation of the states obtaining.'[103] As a result of this, in NP there is an, 'illusion of influence', since the action is referred to in an explanation of the states obtaining:

Why was there $1million in the box? Because the Predictor predicted that you would one-box.

However, it is referred to in a non-extensional belief context, thus not illustrating true influence.

It is for this reason that one may mistakenly believe there to be causal influence in NP, 'it is apparently a persistent temptation for people to believe, when an explanation of something x brings in terms referring to y in a non-extensional belief context... that y, in some way, influences or affects x.'[104] Thus, the agent may come to believe that his

103Nozick (1997) pp67.
104Pp68.

action will cause the Predictor's prediction. Acting accordingly, he would one-box so that $1million was placed in b2. This is in contrast to the questions that would be asked in the other common cause cases, in which the agent's actions are not referred to in the explanation of the states obtaining.

Why did I get cancer? Because you had a gene that gave you cancer.
Why did I get a terrible disease? Because you inherited a gene from your father for the disease.

Why didn't I get the job? Because you failed the test.

There are obvious questions that would need to be asked with regards to such a hypothesis. For instance, this hypothesis seems to rely on NP being described in such a way as to prompt the agent's mistaken belief that his action can affect the Predictor's prediction.

The question 'Why was there $1million in the box?' with its attendant answer 'because the Predictor predicted that you would one-box' is not in fact analogous to the question: 'Why did I get cancer? with its attendant answer 'because you had a gene that gave you cancer', since the gene is the equivalent of the agent's brain structure, while cancer is the equivalent of the prediction of two-boxing. The lack of a description of the common cause in NP makes this question seem to be the appropriate one to ask. If NP were described in its completeness, with a description of the common cause, it would seem that the appropriate question and answer would be:

Why was there a prediction of one boxing? Because of your brain structure.

Thus, empirical evidence would be required to support the idea that NP causes the agent to believe he can have a causal influence over the Predictor's prediction. Furthermore, investigating whether a full description of NP (including a description of the common cause) would stop agents from one-boxing would be a worthwhile project.